Efficient partnership models for energy technology startups enabled by artificial intelligence that hyper-targets ecosystem connectivity (Partner AI)

Pit Rho Corporation

Final Technical Report
December 2019

Period Covered: October 1, 2017 - September 30, 2019

DOE Innovative Pathways Program

DOE FOA Number: DE-FOA-0001703
Award Number: DE-EE0008112

PI: Josh Browne, Ph.D
josh.browne@rho.ai

Awardee: Pit Rho Corporation d/b/a Rho AI
DUNS: 80513458
1049 Main Street, Evanston, WY 82930

Acknowledgement

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Office of Technology Transitions Innovative Pathways Program, Award Number DE-EE0008112.

Executive Summary

PartnerAI is a platform that utilizes the tools of the rapidly advancing field of Artificial Intelligence (AI) to understand and identify the connections between founders and potential partners or investors in the clean energy space. The platform makes use of data science tools, including clustering and NLP libraries, to evaluate scraped text and data, and build a real-time network of potential connections.

The 'PartnerAI' Database contains a graph-based network of connections between organizations and individuals that updates as new organizations and people are identified. This database makes use of a vast amount of publicly available structured and unstructured data, to generate a dataset built specifically for the application of data science models and techniques. The database is modular and scalable to accommodate large amount of yet-to-be collected data, and yet-to-be released data science tools. These data include company descriptions, individual work history, and company assets like patents and grants. The data science models built on these data can significantly improve the speed and quality of connections between technology developers, strategic partners, and investment partners by efficiently characterizing companies and providing this as an embedded service.

The Partner AI platform and dataset can stand-alone as a networking and information tool, and can also be integrated into other software tools. At present, Partner AI is being integrated into a carbon reduction analysis tool intended to estimate emissions reduction potential for early stage ventures (CRANE tool) that will be productized as "CRANE-Ventures."

Summary

This project was focused on taking advantage of the tools of the rapidly advancing field of Artificial Intelligence (AI) to speed the connection process between entrepreneurs and partners/investors for clean energy com investment (and eventual commercialization) to maintain a larger, more interconnected network than a single individual can maintain. By applying AI, we developed a commercial networking product, 'PartnerAI' that significantly improves the speed and quality of connections between technology developers, strategic partners, and investment partners by processing comprehensive partnership data sets and contextual information from the web. The successful implementation of this scalable, real-time

networking platform is an example of the power of AI-driven tools and the techniques used in this project are likely applicable in other fields where networking inefficiencies exist.

Introduction

The clean energy sector remains an industry in dire need of connectivity between ideas and capital. Generally, capital most often finds its way to funding projects, programs, and companies through individual connections and previously established professional relationships. This is generally a sector-agnostic problem, but the clean energy sector feels the full effect of unfunded opportunities given the capital-intensive nature of many companies in the sector. Additionally, the populations of individuals developing clean energy technologies and those with the means to secure funding do not always overlap. This inefficient capital distribution method leaves many worthy companies without the investments they need to execute their vision. In summary, the disparate temporal and spatial limitations inherent to the clean energy sector, both from a technology innovation perspective and investment community, presents an opportunity to massively improve the efficiency of network connectivity in this space.

We also find ourselves at the spearhead of the rapidly advancing field of Artificial Intelligence (AI), including toolsets focusing on natural language processing (NLP) and classification. These highly adopted technology areas can enable high-throughput processing of language that could previously only have been accomplished by humans. Combining these capabilities with this industry-specific need allows for a unique technology solution. We proposed to speed the connection process between entrepreneurs and partners/investors for clean energy commercialization/investment to maintain a larger, more interconnected network than a single individual can maintain.

AI can super-charge the connection process and allow entrepreneurs to find better business partners. Simple exponential mathematics reveals that a growing energy technology startup that makes 12 rate-limiting strategic company-building connections per year and exits following eight years at a 3X multiple to original investors would instead exit at 4.5X or 9.0X if it reduced the timeline by one and two weeks, respectively, between each strategic company-building connection. The cumulative and compounded impact may be transformational in catalyzing venture capital to these innovations providing returns comparable or exceeding those in the software sector.

Applying AI, we worked to develop this commercial networking product, ‘PartnerAI’ that significantly improves the speed and quality of connections between technology developers, strategic partners, and investment partners by processing comprehensive partnership data sets and contextual information from the web.

This project was designed to fully test the thesis within the 1st year and culminate in a commercially-viable product ready for use across the energy industry. The development timeline and milestones were modeled after other end-to-end software products developed by Rho AI. The project arc followed a well-established series of gated stages and allows the technical team to operate in an agile fashion. The early phases focused on identifying and vetting required critical technologies and data sources, followed by development of the first working prototypes of the infrastructure, application stack, and AI models.

Methods, Assumptions, and Procedures

Below is a detailed review of the individual milestones for the project.

Tasks for Milestone 1.3 (months 1-6):

- **Data collection and validation:** Compiled and collated initial data sources to seed underlying database and models. Identified potential data partners and developed plans for ongoing data collection.

- **Enumerated potential data sources and data partners:** Compiled a list of potential data sources; document availability and methods of collection. Seeked out data partners (including NREL, DOE, other grant recipients, and venture capital). Reviewed public databases and financial documents as data sources. Documented known sources, the minimum required combination of data sources, and charted a data path forward for the remainder of the project.

- **Created list of initial data inputs:** Compiled list of preliminary data inputs, representing (a) individual investors, (b) venture capital firms, (c) entrepreneurs, (d) incorporated entities, (e) supply chain companies and corporate strategic partners and (f) spinouts and pre-spinouts from labs and universities that will form the initial data set. Focused on methods to find examples from all six groups to define the ecosystem of organizations defining the space. These companies defined the initial universe of entities that populated the application.

- **Validated external data sources:** Focused on accessing and representing the external data sources internally in the application (e.g. financial reports, website text, and key contacts). The internal representation of the data (i.e. the method and format of storage, specifically the type of database) was a critical step as it is the foundation upon which all data science and AI is built. This work focused on verifying available data sources, which consisted of a technical validation for each external data source used. The validation step evaluated the method to: (a) access the information, (b) parse the information once accessed, and (c) store the data in our internal databases.

Milestone 1.3: Documented technical evaluation of data sources and recommended approaches for accessing the data.

Comments: Upon the completion of Milestone 1.3, we collected and evaluated numerous structured and unstructured data sources, which included data on organizations as well as individuals. The data sources focused on financial news, based on our assumptions that finding and following transactions in the clean energy space would lead us to financial hubs in the space, helping build our graph network. Many of these datasets were publicly accessible datasets or publicly facing websites, including financial news websites, a relatively unique dataset on which we could subsequently build data science models.

Our efforts focused on comprehensiveness, as we strived to find information on all possibly relevant companies in clean energy, sometimes before the organizations had a large online presence.

Tasks for Milestone 2.1 and 2.3 (months 6-12):

- **Development of alpha software stack and data models:** Developed the preliminary data models required to suggest (a) potential investors and strategic partners to entrepreneurs and (b) potential entrepreneurs and investments to investors and strategic partners. In parallel, developed the foundational software and database technologies required to deliver predictions and recommendations to users.

- **Establish critical tech stacks, deploy underlying cloud infrastructure**: Mapped out and documented product requirements, and define critical technologies. Deployed underlying cloud infrastructure to support software. Infrastructure requirements agreed upon and implemented, and critical components of application technology stack and AI software defined.

- **Preliminary model prototypes developed:** Developed prototype models required to (a) suggest potential investors and partners to entrepreneurs and (b) suggest potential entrepreneurs and investments to investors and strategic partners. These models use the data defined and collected in previous steps to generate preliminary models. Key tools utilized were (a) “natural language processing” which enables the use of unstructured text from locations like websites, financial documents, and correspondences as an input to AI models, and (b) Recommender systems, which enable recommendations for users based upon the desires of similar users (e.g. Netflix®). The preliminary models were based upon a subset of companies described in previous tasks.

- **Development of alpha API completed**: Created the first basic software version of the product that is able to interact with the outside world through an application programming interface (API). Ensured that Alpha-level components (data ingestion approach, database design, method for data processing, and web API) were operational.

Milestone 2.1: The cloud network should be accessible with baseline technology stack implemented. A software repository (i.e. “git repository”) will be established, including a requirements.txt file containing all primary application-layer software and dependencies.
Milestone 2.3: A verifiable experimental environment with demonstrable input/output of data will be delivered and data science models will be verifiable through command-line interactions. This will be a development environment, verification of which can be done through live demos, screen shares, or remote computer access.

Comments: These milestones were achieved through the creation of internal tools - Entity Match and Primary Research Tool, primarily by using a heavily modified closed source NLP library programm, “Prodigy.” Additional characterization of our needs led us to choose a graph database, which is discussed in the Results section.

Tasks for Milestone 3.1 (months 12-15):

- **Transitioned alpha software stack and components to beta software release:** Alpha level software is a prototype system that is meant only to test the functionality of the software stack. Beta level

software is considered to be usable and ready for testing, but not yet ready for production level release.

Milestone 3.1: A beta-level back-end demonstrating that the program merits the development of a sophisticated front-end for user interaction and the continued development of the back-end and data science. At this stage, the product will essentially be a Beta API.

Comments: This milestone was achieved through the refinement of our labeling software, the establishment of our data entry team's interface for labeling news that enters our database, and additional tooling to improve speed.

Tasks for Milestone 4.1 (months 12-15):

- **Business model development**: Performed customer discovery to evaluate willingness-to-pay of investors and strategic partners, and incorporated this feedback into a draft concept for sustaining the tool long-term (e.g. open-access base with added priced service layer).

Milestone 4.1: Draft concept to sustain the tool long-term. Produce brief report summarizing customer discovery feedback, willingness-to-pay, and draft concept.

Comments: Our efforts in customer discovery led us to a few conclusions, which we discuss in the results section. We generated a list of top connected investors and program directors in clean energy, and added them to our newsletter distribution as well as our real-time alerts system (Network Aware Real Time Alerts, or "NARTA").

Tasks for Milestone 5.2 (months 12-18):

- **Front-end interface development:** Create a UI (User Interface) to allow potential users to interact with the software system.

- **Alpha level (prototype) front end interface**: Implement an alpha-level front-end interface to interact with the underlying APIs. Success will be measured by delivering an operational web application accessible over the internet. The final product vision is a web-based terminal allowing users to perform customized searches.

- **Beta level (testing) front end interface**: Implement a staging (beta)-level front-end interface. Verification will be based on an operational web application accessible over the internet.

Milestone 5.2: Beta-level front end interface release. The interface would allow a user to interact with the data set as to obtain key events and information about entities being tracked in the product.

Comments: This milestone was achieved with a beta-level front end interface that we displayed at the ARPA-E conference in 2019.

Tasks for Milestone 6.1 (months 12-18):

- **Integrated model development:** Integrated model, back-end software, and front-end software into a production-level software tool.

- **Upgrade model performance**: Reviewed the performance of the model by soliciting feedback from field test users on the quality of recommendations offered, and identify areas for improvement. Based upon feedback, further develop models required to suggest potential investors and strategic partners to entrepreneurs, and suggest potential entrepreneurs and investments to partners and investors. Further model developments may include: (a) new sources of data for more complete coverage, (b) new data types to incorporate previously unconsidered variables, (c) modifications to the model algorithms, and (d) data cleansing.

Milestone 6.1: Create a beta version model that performs functions (a) and (b) above. The model would be validated by a statistical evaluation of the model framework. Existing data sets will be split into training and testing data. The training set will be used for a predictive model to associate investors and technology providers. Accuracy will be assessed via an accuracy measurement known as F1-score. An F1-score of 90% is considered excellent, and 75% or greater is good. For this beta-level model, an F1-score of at least 85% will be targeted. Provide a brief report documenting the results.

Comments: As we discuss in our results section, we modified our approach to focus less on the F1-score as a barometer for success, instead relying on feedback from our real-time alerts and our asynchronous clustering of companies into industries. Therefore, our beta methods and tasks was changed to: continue to refine real-time alerts system and solicit market feedback on tool efficacy at finding the correct news. Our success thresholds remained the same and were hit.

Additionally, with our new goal to refine the real-time alerts, we developed internal automated tools that helped us scrape generic websites in an unstructured format, adding data to our database that was previously inaccessible.

Tasks for Milestone 7.2 (months 18-24):

- **Integrate model into Beta-level software:** Integrated model code into the software stack so that the model can be called from the beta-level user-interface, and returned relevant results.

- **Production Product Release:** Created the production-ready release of the software. Production ready code was ready for operation and public consumption.

- **Production infrastructure development**: Introduced production-level components of the application. High-availability implementations of all critical components were delivered, including: data ingestion, databases, data processing, web APIs, and web services. Verification took place through the validation of high availability through systematic outages and planned failures.

- **Production product release**: This consisted of deploying the first widely useable, stable version of the software to the production infrastructure (servers).

Milestone 7.2 (End project goal): A commercially available PartnerAI Platform that allows users to identify strategic partnership and investment opportunities through an interactive search process, updating in real-time as new data become available.

Comments: Based on user feedback and concurrent partnerships, we added to this milestone an integration with the CRANE Tool, being developed with other channel partners in the Innovative Pathways program to determine emissions reductions potential for small private companies.

Results and Discussion

Results for Milestone 1.3

We determined that our best data sources were funding news sources, such as SEC EDGAR (Form D), Crunchbase, Pitchbook, and other clean energy specific databases. Additionally, we found that unstructured news sources such as generic news websites could give us some insight into funding events that were not reported by traditional funding news sources.

Additionally, we found that organization descriptions were very helpful in our data science efforts to classify and sort companies. Therefore, we extracted data from sources that had descriptive keywords and full paragraph descriptions about companies.

For individuals, we found that work social network accounts were effective in giving us an understood network of connections and related organizations. We augmented that data with data from unstructured sources such as news mentions and profiles on other sites.

Results for Milestones 2.1 and 2.3

Milestones 2.1 and 2.3 focused on the development of the alpha software stack and data models. To complete this task, we chose Amazon Web Services (AWS) as our cloud provider because it provided the most flexibility and extensibility when it comes to deploying applications and services.

Additionally, we developed several custom data scrapers to parse both structured and unstructured data sources such as Crunchbase and Business Wire, among others. To support these services, we deployed a mix of Elastic Compute(EC2) spot instances, a self-hosted graph database (ArangoDB), and managed services such as Elasticsearch to store and process our data.

Further, we chose to use a product called Prodigy to quickly label and annotate scraped data. This allowed us to generate data sets for training new machine learning models. Similarly, Prodigy was used to build our first version of Entity Matching and Research systems. These components were built so we could easily match new entities against our existing data set, and also augment existing entities with new data.

All of the aforementioned services and infrastructure components were deployed using industry standards. Specifically,

- all custom software components are tracked using internal Git repositories and managed using semantic versioning,

- all self-hosted services are configured and deployed using Ansible, Docker and Rancher and their configuration is tracked in Git repositories, and

- all software and infrastructure components have been developed and deployed with high-availability when applicable.

Results for Milestone 3.1

During this milestone, we concentrated on data consolidation, model improvements and development of alpha interface of news and entity exploration.

For data consolidation, we decided to use a service called Event Registry which is an online news aggregation and delivery service from which we can capture news articles from most online providers. Instead of building many scrapers for individual sources, we built one scraper against Event Registry. This resulted in being able to download ~100k documents per month, and create a rich data set for extracting entities and important events.

In parallel, using Prodigy, we quickly developed new NLP models including:

- Easily detection of funding news of key entities being tracked in our graph database.
- Importance of extracted entities
- Identify duplicate events
- Recommendation engine

Further, we built two alpha products based on the data gathered through scraping and entity matching. First, we built a key events generator for tracked entities. These events were quickly distributed through email to our partners. Second, we build a web interface for data exploration. This new web console provided a way for users to explore information about key entities.

Results for Milestone 4.1

The original Statement of Project Objectives focused on creating a recommendation system that would use a F-1 Score to evaluate the efficacy of the project. End-user feedback through our customer discovery process has provided that this evaluation method is not in line with their goals. Instead, the true business driver is in fact real-time alerting related to companies of interest. Ultimately, users are looking for a system that can track all relevant data (people, events, etc) about those companies and keep them informed of important news about these entities. Although users might be interested in new companies that would be uncovered by a recommendation engine, the pain point to be solved is actually just keeping them informed of important news on companies of which they are already aware. In other words, the limitation is not necessarily on knowing what companies are in the universe, but rather keeping track of all those companies. This need really relies primarily on a complete, accurate, well-curated graph database.

Therefore, we built a Network Aware Real-Time Alerts ("NARTA") system that alerted users whenever a funding event occurred in an industry/to a company that was interesting for them as a user. This allowed us to stay engaged with potential long term users, and helped them understand the comprehensiveness of our scrapers. This system consisted of a weekly email update that was personalized for each user, as well as a real-time email alert for highly relevant funding news for select users.

We engaged in conversations with other clients both inside and outside the energy space. The sentiment from these conversations was similar to other customers, in which they want a system that can track all possible information about entities. In addition, one of those engagements unearthed other use cases for PartnerAI, specifically about relationship-building and recruiting. The client highlighted the need to get notified about people that they can engage with, as well as keeping them apprised of any events that can help them hire senior members for their organizations.

Results for Milestone 5.2

During this milestone, we released the beta version of our web console which was showcased at ARPA-E 2019. This release incorporated customer feedback and provided a more cohesive way of exploring important events from key entities, as well as more detailed information about these entities. Currently, any user can explore the entirety of data set, which is composed of ~10k organizations and ~62k people. Additionally, this data set is being used to power other key products at Rho AI as explained further in this document.

Results for Milestone 6.1

As discussed in Results for Milestone 4.1, we determined we would not use the F1-score as a metric for evaluation of our system. Instead, we relied on customer feedback concerning our real time alerts hit rate to guide our progress on scraping and discovering new companies.

As it was mentioned before, we originally used Prodigy for building our Entity Matching and Primary Research Tool. Although these solutions worked at small scale, they proved to be a bottleneck when processing large data sets. Therefore, we chose to rewrite these tools to improve speed and concurrency of data processing. With these new tools we can:

- Process multiple entities in parallel
- Continuously generate new tasks without system restarts
- Improve accuracy of our data

Results for Milestone 7.2

Through this program, we were able to integrate our system with the in-development CRANE Tool, providing key data and information on companies in clean energy for the purposes of understanding emissions reductions potential. Our database, organized in a graphical manner, continues to be a useful resource.

The tools we developed during the course of this program allowed us to generally ingest large amounts of data, clean and categorize the data, and deliver data science results such as clustering of companies and intelligent recommendations. We have leveraged these tools to target external customers who may have overlapping datasets and needs that are beyond their capability to analyze the datasets; our tools have helped immensely.

Conclusions

Throughout the 2-year development cycle of this innovative program, we were successfully able to demonstrate how AI-based tools and techniques can be used to build a real-time, scalable networking tool to connect clean energy innovators with investors and partners. Given the rapidly changing and growing nature of the AI field in general, we found ourselves applying AI tools that weren't necessarily mature enough during our initial scoping of this project, but became more readily available during the project development period. The evolving nature of our deliverables as highlighted in the quarterly reports captures this dynamic in the AI space.

In addition to achieving our project objectives as outlined in detail in this report, we have since been able to incorporate the PartnerAI platform into a customer facing, web-based tool suite known as "CRANE." CRANE (Carbon Reduction Assessment for New Enterprises), funded through grants from Prime Coalition, estimates emissions reduction potential for early stage clean energy technologies. This aligns well with the intended user base for PartnerAI, making a collaboration between Rho AI and Prime to develop CRANE, and incorporate PartnerAI into CRANE, as an efficient next step in the evolution of the PartnerAI platform.